Study on the English Corresponding Unit of Chinese Clause

Wenhe Feng[1,2(✉)], Yi Yang[3], Yancui Li[4], Xia Li[2], and Han Ren[2(✉)]

[1] Wuhan University, Wuhan 430073, Hubei, China
wenhefeng@gmail.com
[2] Guangdong University of Foreign Studies,
Guangzhou 510006, Guangdong, China
200211025@oamail.gdufs.edu.cn, softrl@126.com
[3] Ural Federal University, Yekaterinburg 620000, Russia
xwyang@mail.ru
[4] Henan Institute of Science and Technology, Xinxiang 453003, Henan, China
yancuili@gmail.com

Abstract. This paper annotates the English corresponding units of Chinese clauses in Chinese-English translation and statistically analyzes them. Firstly, based on Chinese clause segmentation, we segment English target text into corresponding units (clause) to get a Chinese-to-English clause-aligned parallel corpus. Then, we annotate the grammatical properties of the English corresponding clauses in the corpus. Finally, we find the distribution characteristics of grammatical properties of English corresponding clauses by statistically analyzing the annotated corpus: there are more clauses (1631,74.41%) than sentences (561,25.59%); there are more major clauses (1719,78.42%) than subordinate clauses (473,21.58%); there are more adverbial clauses (392,82.88%) than attributive clauses (81,17.12%) and more non-defining clauses (358,75.69%) than restrictive relative clauses (115,24.31%) in subordinate clauses; and there are more simple clauses (1142,52.1%) than coordinate clauses (1050,47.9%).

Keywords: Clauses · Parallel corpus · Clause-based · Clause alignment · Discourse-based translation · Chinese-to-English translation

Clause is the basic unit of discourse translation. Previous research has shown that in Machine Translation Systems, the acceptance rate of clause-based translation is 45% higher than the sentence-based translation [1]. Thus, clause-based translation model has become an important subject for discourse-based machine translation studies [2]. Nowadays, statistical translation is grounded on the large scale bilingual aligned samples and bilingual grammatical knowledge. Therefore, one of the most important issues for discourse-based machine translation studies appears to be building clause-aligned and -annotated parallel corpora.

In Chinese-to-English translation, English corresponding units of Chinese clauses (ECUCC) in the translation are complex and diverse. One key issue of translation is to select the proper form of ECUCC for translation. For example, in (1), Chinese clauses C1 and C2 are translated to English corresponding units E1 and E2. Functionally, E1 is the major clause and E2 is the subordinate clause (the adverbial clause); structurally, E1 is the restrictive relative clause and E2 is the non-defining clause (the present participle).

C.-Y. Lin et al. (Eds.): NLPCC-ICCPOL 2016, LNAI 10102, pp. 129–140, 2016.
DOI: 10.1007/978-3-319-50496-4_11

(1) [C1]上海浦东近年来颁布实行了涉及经济、贸易、建设、规划、科技、文教等领域的七十一件法规性文件,/[C2]确保了浦东开发的有序进行。

[E1]In recent years Shanghai's Pudong has promulgated and implemented 71 regulatory documents relating to areas such as economics, trade, construction, planning, science and technology, culture and education, etc., /[E2]ensuring the orderly advancement of Pudong's development.

As shown in (1), if we can provide bilingual clause-aligned samples and add grammatical annotations for clauses, it will not only provide an effective guide for the translation of parallel clauses, but will also lay a foundation for discourse-based Machine Translation Studies. At present, several Chinese-English parallel corpora have been built on the sentence- (usually marked with period) or paragraph-alignment [3, 4]. Studies on clause-aligned parallel corpus at a preliminary stage [2] that there are few annotated resources of grammatical knowledge for segmenting parallel texts into clauses.

In this paper, we recount our experience in annotating ECUCC and statistically analyze them. Firstly, based on Chinese clause segmentation we segment English texts into corresponding units for parallel text to get a Chinese-to-English clause-aligned parallel corpus (Sect. 1). Then, we annotate the grammatical s of the ECUCC in the corpus to get a grammatical annotated corpus (Sect. 2). Finally, based on the annotated corpus we find the distribution characteristics of grammatical properties of ECUCC by statistically analyzing the annotated corpus (Sect. 3).

1 Chinese-to-English Clause-Aligned Parallel Corpus

Building a Chinese-to-English clause-aligned parallel corpus is based on the following principles: (1) define the rules for Chinese clause segmentation in Chinese-English parallel texts; (2) based on the results of Chinese clause segmentation, divide English translated texts into units, and get the best English corresponding units in a linear sequence, which are ECUCC.

The rules for Chinese clause segmentation in our study applies the definition of clause by Li [6, 7]: "Clause is the basic unit of discourse analysis, including the traditional simple sentences and clauses in compound sentences. Structurally, an independent clause contains at least one predicate and at least one proposition; functionally, an independent clause is not used as any grammatical component to other clauses, and there is only propositional relationship between two independent clauses; formally, there must be punctuation (comma, semicolon, or period) between two independent clauses. Besides, some traditional phrases, which are similar to typical clauses in structure, function and forms are treated as clauses." Studies [5, 6] have shown that such definition of Chinese clause provides operability to create and automatic analyze large-scale annotated corpus.

The "based on the results of Chinese clause segmentation divide English translated texts into units" means that we divide English translated texts based on the results of segmentation Chinese clauses. In example (2), Chinese text is divided into three clauses which are marked as C1, C2 and C3 and accordingly, English corresponding unit are

divided as E1, E2 and E3. Grammatically, E1 is a typical clause, E2 and E3 are not. E2 is a clause group, and E3 is an infinitive phrase. According to the nature of English, E2 would be divided into two English clauses ("…expand…" and "use…"). But we analyze E2 which is the corresponding unit to Chinese clause C2 as the final unit, based on the rules for Chinese clause segmentation. Therefore, we call E1, E2 and E3 as English corresponding unit of C1, C2 and C3. However, sometimes we also call these corresponding units of Chinese clause as "English clause".

(2) [C1]浙江省今后将进一步提高对外开放水平，//[C2]努力扩大对外贸易、利用外资和国际经济技术合作，/[C3]并逐步完善对外经贸营销网络。

[E1]Zhejiang Province will further raise the level of opening up to the outside world, //[E2]diligently expand its foreign trade, and use foreign funds and international economic and technical co-operation, /[E3]to progressively perfect its marketing network of foreign economic and trade business.

(3) [C1]这一数字比上年末增加二百零三亿元，/[C2]增长百分之二十七。

[E1]This number was an increase of 20.3 billion yuan, /[E2]a growth of 27% compared to the end of the previous year.

The "best English corresponding units in a linear sequence" means that the English corresponding unit segmentation should correspond to the Chinese clauses in a linear sequence, but not necessary in semantics. For example, in (3), E1 and E2 semantically are not equal to the C1 and C2 because of the position of the adverb (compared to the end of the previous year). In this case, E1 and E2 are the best English corresponding units in a linear sequence of C1 and C2.

Based on the above principles, we select 100 Chinese-English parallel texts (news) to build a Chinese-to-English clause-aligned parallel corpus, in which Chinese clauses and their English corresponding units are aligned.

2 ECUCC Grammatically Annotated Corpus

In the ECUCC grammatically-annotated corpus, 2192 ECUCC taken from the Chinese-to-English Clause-Aligned Corpus are analyzed and annotated. Grammatical properties of ECUCC are analyzed and annotated under certain principles and systems.

2.1 Grammatical Analytic Principles of ECUCC

To deal with problems of grammatical analysis of English corresponding units, we formulated the analytic principles through analysis and verification.

First, in the process of identification of the grammatical properties of ECUCC, both their inner structure and external function should be considered. As shown in Example (1), structurally, the core verbs in E1 and E2 are different between restrictive relative and non-defining; functionally, general structures are different between the major and subordinate in the global structure.

Second, for identifying the major object of ECUCC, the global function takes priority over the local function. Sometimes ECUCC is complicated in the inner structure, and it is difficult to identify its grammatical properties. In this case, the identification of the structure and function is based on the major object of the unit, while the identification of the major object is based on the global function of its global structure. For example, in (4), E1 is complicated by its inner structure (it consists of major clause and adverbial clause, while adverbial clause is composed of coordinate attributive clauses). The whole sentence is a complex sentence: E1 is the major clause, E2 is a subordinate clause. Thus, E1 can be identified as "major clause + finite structure" according to the function of major object ("recently there were…").

(4) [C1]据浦东新区经贸局对浦东开发七年来引进投资一千万美元以上的一百五十七个工业大项目跟踪调查, 目前建成投产的有一百一十六个, /[C2]投产率高达百分之七十三点九。

[E1]According to the Pudong New Region's Economy and Trade Bureau follow - up investigation into 157 large industrial projects that were introduced in the seven years of Pudong's development, and that have more than 10 million US dollars invested, recently there were 116 that finished construction and went into operation, /[E2] with the percentage of going into operation reaching up to 73.9%.

Third, sometimes omissions in ECUCC influence the identification of their grammatical properties. In this case, the analysis should be based on the completed sentence. For example, in (5), there is an ellipsis of preposition "with" in clause E3 and E4. It is required to complete E3 and E4 before the analysis. Thus, E3 and E4 are identified as "coordinate" "prepositional phrase" and "adverbial".

(5) [C1]去年一至十一月, 内地在香港新签对外承包工程、劳务合作和设计咨询合同一千四百七十四份, /[C2]合同金额二十点九四亿美元, //[C3]完成营业额十五点八亿美元, //[C4]输港派出劳务二万一千一百五十三人次。

[E1]From January to November of last year, the inland signed 1,474 new contracts for foreign contracted projects and cooperation of labor service and design consultation in Hong Kong, /[E2]with a contracted value of 2.094 billion US dollars, //[E3]a completed turnover of 1.58 billion US dollars //[E4]and 21,153 man - times of labor service sent to Hong Kong.

2.2 Grammatical Analytic System of ECUCC

Based on the studies of the corpus, the grammatical analytic system has been functionally and structurally formed [7] (Further details follow in Sect. 3).

Functionally: firstly, according to the grammatical properties of a whole sentence (simple sentence, coordinate sentence, complex sentence) and the position of a clause, English clauses can be divided into independent clauses, coordinate clause, major clauses and subordinate clauses; secondly, according to the function, clauses can be divided into adverbial clauses, attributive clauses and so on; finally, according to quantity of clauses with the same function in a sentence, clauses can be divided into simple clauses and coordinate clauses.

Structurally: firstly, according to the properties of predicate verbs, clauses can be divided into restrictive relative clauses and non-defining clauses; secondly, depending on particular conditions, non-defining clauses can be divided into infinitive, present participle, past participle, non-verb, preposition structure and other subcategories.

3 Classification and Statistical Analysis of ECUCC

3.1 Sentences and Clauses

ECUCC may be a sentence, or a clause. Separate sentence as example (6) and clause group as example (7) can independently performed an utterance function. Clauses which include coordinate clauses and various types of major or subordinate clauses (see Sect. 3.2) cannot performed an utterance function. It should be combined with other clauses to form a complete sentence.

(6) [C1]建筑是开发浦东的一项主要经济活动，/[C2]这些年有数百家建筑公司、四千余个建筑工地遍布在这片热土上。

[E1] Construction is a principal economic activity in developing Pudong. /[E2]These years there have been several hundred construction companies and over four thousand construction sites that have spread out all over this stretch of hot turf.

(7) [C1]在世界经济一体化与日俱增的环境下，各国面对全球化带来的挑战，应通过持续推行健全的经济政策以及深化结构改革来从全球化进程中最大限度地受益并把负面影响减少到最小程度。

[E1]The unification of the world economy is intensifying with each passing day. Facing the Challenges brought by globalization, each country should continuously implement sound economic policies and deepen structural reform so as to enjoy the most benefits from the process of globalization and to minimize the negative effects.

The statistical distribution of sentences and clauses of ECUCC is given in Fig. 1. The results show that clauses are more than sentences by three times which indicates that Chinese clauses are more likely to be translated as English clauses rather than English sentences.

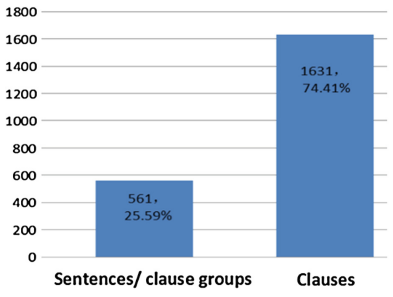

Fig. 1. Statistical distribution of sentences and clauses in English corresponding units

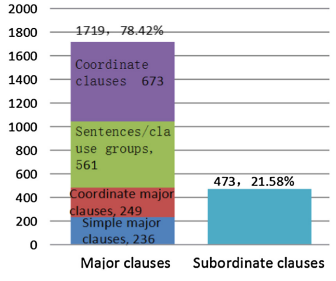

Fig. 2. Statistical distribution of major clauses and subordinate clauses in English corresponding units

3.2 Major Clauses and Subordinate Clauses

ECUCC may be characterized by major clauses or subordinate clauses. English major clause units include simple major clauses (example 9) and coordinate major clauses (example 10) (details of simple/coordinate clauses follow in Sect. 3.5 below), coordinate clauses (example 8), sentences (example 6) and clause groups (example 7). The major clause units are generally finite structures and can be independently used as sentences. Subordinate clause units include 20 kinds of clauses such as attributive clauses, adverbial clauses, infinitive, and present participle clauses (see Sects. 3.3, 3.4 and Table 1). Subordinate clause units are barely used as independent sentences.

(8) [C1]去年十月，中国进出口银行聘请日本野村证券公司作顾问，/[C2]向日本著名的评级机构日本公社债研究所提出正式评级申请。

[E1]LastOctober, the Import and Export Bank of China invited Nomura Securities of Japan to be advisors, /[E2]and submitted a formal assessment application to the Commune Bond Research Institute of Japan, a famous assessment institution in Japan (Coordinate clause).

(9) …… [C1]而是借鉴发达国家和深圳等特区的经验教训，/[C2]聘请国内外有关专家学者，//[C3]积极、及时地制定和推出法规性文件，//[C4]使这些经济活动一出现就被纳入法制轨道。

……. [E1]Instead, Pudong is taking advantage of the lessons from experience of developed countries and special regions such as Shenzhen /[E2]by hiring appropriate domestic and foreign specialists and scholars, //[E3]by actively and promptly formulating and issuing regulatory documents, //[E4]and by ensuring that these economic activities are incorporated into the sphere of influence of the legal system as soon as they appear (Simple major clause).

(10) [C1]当前经济的关键不是争取更高的增长速度 //[C2] 而是调整结构,提高效益，/[C3]以使一九九三年下半年以来实行的宏观调控取得更大成果，//[C4]把国民经济推上一条持续、快速、健康发展之路。

[E1]The key of the current economy is not striving for a higher growth rate, //[E2]but is adjusting structures and increasing benefits, /[E3]so as to make macro controls which were implemented from the second half year of 1993 obtain greater achievements //[E4]and push the national economy onto a road of constant, rapid and healthy development (Coordinate major clause).

The statistical distribution of major clauses and subordinate clauses in English corresponding units is given in Fig. 2. The results show that major clauses are more than subordinate clauses by four times. It indicates that Chinese clauses are more likely to be translated as English major clauses rather than subordinate clauses.

3.3 Functions of Subordinate Clauses: Adverbial and Attributive

Functions of subordinate ECUCC can be adverbial (examples 11–12) and attributive (examples 13–14).

(11) [C1]如果亚洲的经济形势恶化或者金融危机对外界的影响增大，/[C2]全球原油需求量的增长幅度可能会进一步缩小。

[E1]If the Asian economic situation deteriorates or the outside influence of the financial crisis becomes larger, /[E2]the growth rate of worldwide demand for crude oil may possibly further decrease (Simple adverbial clause).

(12) [C1]由于茅台酒制作工艺复杂，/[C2]生产周期长，/[C3]因而其产量十分有限。

[E1]Because the art of manufacturing Mao - tai is complicated //[E2]and its production cycle is long, /[E3]the output of Mao - tai is extremely limited (Coordinate attributive clause).

(13) [C1]中国进出口银行最近在日本取得债券信用等级AA –，/[C2]这是日本金融市场当前对中国银行的最高债券评级。

[E1]Recently, the Import and Export Bank of China won a bond credit rating of AA - in Japan, /[E2]which is currently the highest bond rating given to a Chinese bank by the Japanese financial market (Simple attributive clause).

(14) [C1]据统计，在目前已投产外资大企业的主要产品中，有一百零二个品牌，/[C2]其中国外品牌五十二个，//[C3]国内品牌五十个。

[E1]According to statistics, among the main products of large foreign funded enterprises that have currently been put into production, there are 102 brands, /[E2]of which 52 are foreign brands //[E3]and 50 are domestic brands (Coordinate attributive clause).

The statistical distribution of adverbial clauses and attributive clauses in English corresponding subordinate clauses is given in Fig. 3. The results show that adverbial clauses are more than attributive clauses by five times. It indicates that English corresponding subordinate clauses are translated as adverbial clauses in most situations.

3.4 Structures of Subordinate Clauses: Restrictive Relative and Non-defining

Depending on core verbs, English subordinate clauses can be divided into restrictive relative clauses and non-defining clauses. Core verbs in restrictive relative clauses vary in terms of tense (for examples 11–14), core verbs in non-defining clauses not vary in terms of tense or omitted. Non-defining verbs can be divided into infinitive (example 15), present participle (example 16), past participle (example 17), non-verb (example 18), nominative absolute structure (example 19), prepositional phrase (example 20) and other structural forms.

(15) [C1]进出口银行决定先在日本取得信用评级是为进入国际资本市场融资创造作准备，/[C2]以便扩大资金来源。

[E1]The reason behind the decision by the Import and Export Bank of China to obtain a credit rating in Japan first is to prepare for entry into the international capital market for financing, /[E2]so as to expand sources of funds (Infinitive).

(16) [C1]据统计，目前在纽约证交所上市的外国企业已达 340 多家，/[C2]为5年前的三倍。

[E1]According to statistics, currently, foreign enterprises listed on the New York Stock Exchange have reached more than 340, /[E2]tripling the Fig. 5 years ago (Present participle).

(17) [C1]在经营方面, [C2]该行加强了存款工作, /[C3]使人民币存款的增幅回升, /[C4]同时通过签订银企合作协议和加强对大客户服务等方式, 发展有潜力的优质客户。

[E1]Regarding operations, this bank strengthened deposit work, /[E2]made RMB deposit growth rate come back, /[E3]at the same time, through methods such as signing bank - enterprise cooperation agreements and strengthening services to major clients, etc., [E4]developed potential high grade clients (Past participle).

(18) [C1]东亚首脑非正式会晤在历史上尚属首次, [C2]这是一个良好的开端。

[E1]This informal meeting of heads of Eastern Asian countries, the first time in history, [E2]is a good start (Non-verb).

(19) [C1]报告说, 1997年是经济转轨国家自停止实行中央计划经济以来的第一个经济增长年份, /[C2]增长率达百分之一点七, /[C3]1998年预计增长百分之三点二五。

[E1]The report said that 1997 was the first year of economic growth for those countries with transitioning economies since they had stopped implementing centrally planned economies, [E2]the rate reaching 1.7%, [E3] and estimated to grow by 3.25% for 1998 (Nominative absolute structure).

(20) [C1]镍被称作"现代工业的维生素", /[C2]其合金有三千多种, /[C3]是发展航天、航空、军事和现代科技的特需材料。

[E1]Nickel, called the "vitamin of modern industry", [E2]and with more than 3,000 varieties of alloy, [E3] it is the material specially required to develop space - flight, aviation, military and modern science and technology (Prepositional phrase).

The statistical distribution of restrictive relative clauses and non-defining clauses in English corresponding subordinate clauses is given in Fig. 4. The results show that:

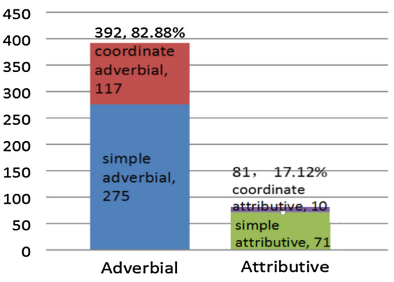

Fig. 3. Statistical distribution of adverbial clauses and attributive clauses in English corresponding subordinate clauses

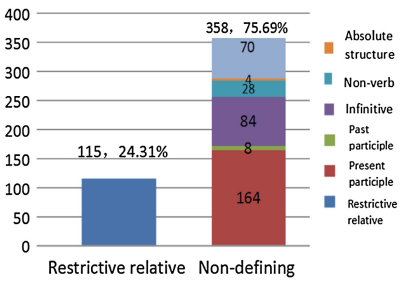

Fig. 4. Statistical distribution of restrictive relative clauses and non-defining clauses in English corresponding subordinate clauses

(1) There are more non-defining clauses (358, 75.69%) than restrictive relative clauses (115, 24.31%). It indicates that English corresponding subordinate clauses are translated as non-defining clauses in most cases.
(2) In the non-defining clauses, the above three categories (present participle structure, infinitive and prepositional phrase) account for nearly 90% of the total. The other three structures (non-verb, past participle, nominative absolute structure) account for 11% of the total. It indicates that English non-defining clauses are more likely to be translated as present participle structures, infinitive structures and prepositional phrase structures than others.

3.5 Simple Clauses and Coordinate Clauses

English clauses are divided into simple clauses and coordinate clauses according to their function. In coordinate clauses, two or two more English clauses perform the same function. Simple clauses can be divided into sentences (example 8), simple major clauses (example 9), simple adverbial subordinate clauses (example 11) and simple attributive subordinate clauses (example 13). Coordinate clauses can be divided into coordinate clauses (example 8), coordinate major clauses (example 10), coordinate adverbial subordinate clauses (example 12) and coordinate attributive subordinate clauses (example 14).

The statistical results show that: (1) simple clauses (1142, 52.1%) are slightly more than coordinate clauses (1050, 47.9%) (Fig. 5); (2) coordinate major clauses (923, 53.69%) are more than simple major clauses (796, 46.31%) (Fig. 6); (3) simple subordinate clauses (346, 73.15%) are much more than coordinate subordinate clauses (127, 26.85%) (Fig. 7).

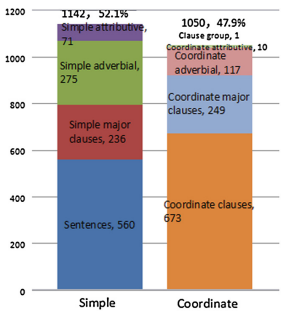

Fig. 5. Distribution of simple/coordinate clauses

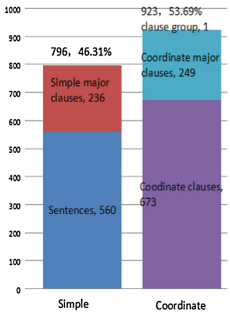

Fig. 6. Distribution of simple/coordinate major clauses

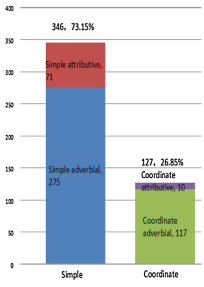

Fig. 7. Distribution of simple/coordinate subordinate clauses

3.6 General Analysis

1. Distribution of Types of ECUCC

The Table 1 summarizes the distribution of types of ECUCC. ECUCC are grouped in the table by frequency range (high-frequency, intermediate frequency and low-frequency). The table shows that: (1) There are 4 types of ECUCC of high-frequency (X > 10%) in the corpus which account for 78.38% of the total distribution. Compared with other grammatical types (except clause group) these four types clauses are major clause units (see Table 2). (2) In the corpus 8 types of ECUCC of intermediate frequency (1% < X < 10%) account for 18.93% of the total distribution, which including 7 adverbial subordinate clauses (16.01%) and 1 attributive subordinate clause 2.92%. (3) 13 types of ECUCC of low-frequency (X < 1%) account for 2.69% in the corpus. In addition to clause group, the remaining 12 categories are subordinate clause units.

Table 1. Distribution of types of ECUCC

Frequency range	Types	No.	%	Frequency range	Types	No.	%
X > 10%	Coordinate clause	673	30.70	X < 1%	Simple adverbial - non-verb	21	0.96
	Independent clauses	560	25.55		Coordinate attributive - restrictive relative	8	0.36
	Coordinate major clause	249	11.36		Coordinate attributive - restrictive relative	6	0.27
	Simple major clause	236	10.77		Simple adverbial - past participle	5	0.23
1% < X < 10%	Simple adverbial - present participle	121	5.52		Coordinate attributive - non-verb	5	0.23
	Simple adverbial - restrictive relative	64	2.92		Coordinate attributive - -ing	3	0.14
	Simple adverbial - infinitive	46	2.10		Simple attributive - present participle	3	0.14
	Simple adverbial - prepositional phrase	44	2.01		Simple attributive - past participle	2	0.09
	Coordinate adverbial - present participle	40	1.82		Simple attributive - non-verb	2	0.09
	Coordinate adverbial - infinitive	38	1.73		Simple adverbial - -ing	1	0.05
	Simple adverbial - restrictive relative	37	1.69		Coordinate attributive - past participle	1	0.05
	Coordinate adverbial - prepositional phrase	25	1.14		Coordinate attributive - prepositional phrase	1	0.05
					Clause group	1	0.05
Total					2192		100.00

2. Distribution of Grammatical Functions of English Corresponding Units

The distribution of grammatical functions of English corresponding units is given in Table 2. In general, there are two features: (1) In terms of quantity, there are 1719 major clause units (78%) and 473 subordinate clause units (22%). The former is about 4 times of the latter. Thus the major clause units are more important in Chinese-to-English translation. (2) In terms of structure and function, the major clause units are more complex than subordinate clause units. Structurally, the core verbs in the major clause units are usually finite verbs, but in subordinate clause units there are different forms of core verb such as: infinitive, present participle and so on; functionally, all types of major clause units can be independently used as sentences. Subordinate clauses are different between adverbial, attributive and others. Therefore, the difficulty of Chinese-to-English translation is the translation of the subordinate clause units.

Table 2. Distribution of grammatical functions of English corresponding units

Functional structure	Major clause units				Subordinate clause units				Total
	Independent clauses	Coordinate	In complex clauses		Adverbial		Attributive		
	Simple	Coordinate	Simple	Coordinate	Simple	Coordinate	Simple	Coordinate	
Restrictive relative	560	673	236	249	37	6	64	8	1833
Present participle					121	40	3		164
Infinitive					46	38			84
Prepositional phrase					44	25		1	70
Non-verb					21	5	2		28
Past participle					5		2	1	8
Independent structure					1	3			4
Clause group	1								1
Total	561	673	236	249	275	117	71	10	2192
	1719				473				

4 Conclusion and Further Research

In this paper, we annotate and present the grammatical properties of ECUCC in the Chinese-to-English clause-aligned parallel corpus. It is of a great significance to Chinese-to-English translation. However, it should be noted here that:

(1) Chinese-to-English translation is different from English-to-Chinese translation. It is necessary to distinguish the two translation directions during the analyzing process. The next step of our work is to build an English-to-Chinese clauses-aligned corpus. The basic idea is the same as building the Chinese-to-English clause-aligned and -annotated parallel corpus.

(2) It is still unknown the grammatical properties of Chinese clauses in the source texts due to the lack of annotations. Therefore, in the future work, grammatical properties of the Chinese clauses also will be annotated. Another paper illustrating the problem of Chinese clauses will be written.
(3) Building the Chinese-to-English clause-aligned and -annotated parallel corpus is grounded in the theoretical framework of Chinese-English discourse structure parallel corpus [8]. The grammatical annotation of ECUCC is one of the important problems under the perspective of discourse structure. In the following works, our studies will improve and expand the scale of both corpora.

Acknowledgments. This paper was supported by Program of humanities and Social Sciences of Ministry of Education (13YJC740022, 15YJC740021), Major projects of basic researches of Philosophy and Sociology in colleges, Henan (2015-JCZD-022), China Postdoctoral Fund (2013M540594), National Natural Science Foundation of China (61273320, 61502149, 61402119), China Scholarship Council (201508090048) and Programs to Improve Competitiveness, Russia (02.A03.21.0006).

References

1. Wang, J.: Computer-Oriented Chinese Translation Studies of English Clauses. Beijing Language and Culture University Press, Beijing (2009)
2. Song, R., Ge, S.: English-Chinese translation unit and translation model for discourse-based machine translation. J. Chin. Inf. Process. **29**(15), 125–135 (2013)
3. Bai, X., Chang, B., Zhan, W., Wu, Y.: The construction of a large-scale Chinese-English parallel corpus. In: Proceeding of 2002 National Machine Translation Conference on Advances in Machine Translation Studies (2002)
4. Wang, K.: Bilingual Corpus: Development and Application. Foreign Language Teaching and Research Press, Beijing (2004)
5. Li, Y., Feng, W., Zhou, G., Zhu, K.: Research of Chinese clause identification based on comma. Acta Sci. Nat. Univ. Pekin. **49**(1), 7–14 (2013)
6. Li, Y., Feng, W., Sun, J., Kong, F., Zhou, G.: Building Chinese discourse corpus with connective-driven dependency tree structure. In: Proceedings of EMNLP, pp. 2105–2114 (2014)
7. Zhang, Z.: A New English Grammar Coursebook. Shanghai Foreign Language Education Press, Shanghai (2013)
8. Feng, W.: Alignment and annotation of Chinese-English discourse structure. J. Chin. Inf. Process. **27**(6), 158–164 (2013)